# ISMT E-117 FINAL PROJECT FALL 2019

# TOPIC MODELING OF NEWSGROUP POSTS

## Team Members:

Cecilia Butler, Cole Crawford, Nikhar Shah, Gokhan Uluderya

## INTRODUCTION

In today’s world, we collect a wide variety of text data in large amounts. Extracting information and insights from this data automatically is not just important from an academic perspective but is a need in many areas and businesses. As the size of the text data we collect grows exponentially, we need more advanced techniques to organize, search, and understand that text data.

One such area where grouping of text data, archival, and searching of it is critical is the news and digital media sector. Articles, stories, comments, and posts get categorized and grouped based on their topics, and sub-topics. Users and professionals can then search and access this this information

## PROJECT DESCRIPTION AND GOALS

In this project, we are applying topic modeling to newsgroup posts in order to categorize them based on their content. Our goal is to mimic the true-life scenario we described in the previous section using a curated dataset from UC Irvine.

In order to achieve this, we will

* Exercise NLP techniques to prepare the free-form newsgroup posts for text analytics
* Use various unsupervised and supervised methods to categorize them into various topics
* Categorize them based on not only primary topics but sub-topics
* Compare different approaches and learn and discuss pros and cons

## METHODOLOGY

### Data Set

Description of dataset including summary stats and shape

## Data Preparation

Describe what we did to prepare the data for modeling

### Applying unsupervised learning to perform topic modeling (LDA)

we'd want to understand if LDA would get close to the human-perceived topics

## Used supervised learning to perform topic modeling

look at the accuracy of that based on a test set and compare with LDA results

### Comparing supervised and unsupervised approaches

pick up random articles and test both LDA and unsupervised technique

## SUMMARY OF FINDINGS AND LEARNINGS

* Do LDA topics resemble the labeled topics for the newsgroups?
* What kind of subtopics are we able to find using LDA
* What kind of accuracy are we getting with supervised learning?
  + We need to use an algo to get probabilities
* When we feed LDA or supervised algo the same text, which seem to find the topic better?

## CONCLUSION